Optim Lett (2017) 11:1293–1302
DOI 10.1007/s11590-017-1108-y

ORIGINAL PAPER

Calculating approximation guarantees for partial set cover of pairs

Peter Damaschke[1]

Received: 5 April 2016 / Accepted: 6 January 2017 / Published online: 12 January 2017

Abstract As a part of a heuristic for the fast detection of new word combinations in text streams, we consider the NP-hard PARTIAL SET COVER OF PAIRS problem. There we wish to cover a maximum number of pairs of elements by a prescribed number of sets from a given set family. While the approximation ratio of the greedy algorithm for the classic PARTIAL SET COVER problem is completely understood, the same question for covering of pairs is intrinsically more complicated, since the pairs insert some graph-theoretic structure. The best approximation guarantee for the first greedy step can be rephrased as a problem in extremal combinatorics: Assume that we may place a fixed number of subsets of fixed and equal size in a set, how many different pairs of elements can we cover? In this paper we introduce a method to calculate optimal approximation guarantees, and we demonstrate its use on the smallest set families.

Keywords Partial set cover · Greedy approximation · Extremal set family · Novelty detection

1 Introduction

We say that a set B *covers* every pair $\{u, v\}$ of its elements $u, v \in B$. A family $\mathcal{F}$ of sets covers a pair $\{u, v\}$ if some $B \in \mathcal{F}$ covers $\{u, v\}$. We are concerned with the following problem that we call PARTIAL SET COVER OF PAIRS: Given m subsets $C_1, \ldots, C_m$ of a set C, and an integer $r < m$, select a family of r of these sets C_i that covers a maximum number of *pairs of elements* of C.

✉ Peter Damaschke
ptr@chalmers.se

[1] Department of Computer Science and Engineering, Chalmers University, 41296 Göteborg, Sweden

The more established PARTIAL SET COVER problem asks to cover *elements* rather than pairs. To see that our problem is a special case of it, consider the pairs in C as elements of the set $\binom{C}{2} = \{\{u, v\} | \, u, v \in C, \ u \neq v\}$ and replace every C_i with $\binom{C_i}{2}$. As we observed in [3], PARTIAL SET COVER OF PAIRS is NP-complete and also W [2]-complete with parameter r. Thus we want fast approximate solutions, even for fixed small r. The *greedy algorithm* for PARTIAL SET COVER (OF PAIRS) sequentially takes s sets each of which covers a maximum number of further elements (pairs). For PARTIAL SET COVER, they cover at least a $1 - (1 - 1/r)^s$ fraction of the optimal number of elements covered by r sets, and this is the best possible guarantee for the greedy rule [5,8]. Now it is natural to ask how to obtain better approximation ratios for the special case of PARTIAL SET COVER OF PAIRS.

Unlike partial cover problems, the original SET COVER problem asks to cover *all* elements by a minimum number of sets. Its approximability [1,6] and parameterized complexity (see, e.g., [10]) are well known. Aspects of partial covering are much less studied, as in [4,9]. Covering of pairs is also related to covering of edges by cliques in graphs [2,7], but it is not the same problem.

Our interest in PARTIAL SET COVER OF PAIRS originates from the fast detection of new combinations of items in data streams, e.g., combinations of words in streams of texts about a common topic. We refer to [3] for more details. There, C is the set of different words in a text, $n = |C|$, and $C_1, \ldots, C_m$ are sets of words in some earlier texts with large overlaps $C \cap C_i$ of word content. One can easily enumerate the *new* pairs of elements (words), i.e., pairs that appear in C but not in any earlier C_i, in $O(n^2)$ time. But if a handful of the sets C_i, say r of them, cover already the vast majority of pairs in C, we can exclude these pairs (trivially they are not new) and check the remaining pairs for being new, in subquadratic time. Texts about a common topic are likely to have such large overlaps. It is not an option to simply list the pairs explicitly, as this would already take $O(n^2)$ time. But by a succinct description of the pairs covered by a set family and random sampling strategies, we can do s greedy steps faster. Actually the time can be exponential in s resp. r [3], but this is not an issue, since these parameters are fixed and small.

1.1 Contributions

For the greedy algorithm for PARTIAL SET COVER OF PAIRS we ask: If r sets C_i cover a fraction c of pairs in C, what fraction g is covered by the first s greedy sets? The approach in [3] works only for $c = 1$ and yields coarse lower bounds on g for $c < 1$. Here we present a framework and method to obtain optimal g directly and for general c. The main ingredients are: an asymptotic notion of coverage that yields disjunctions of linear inequalities as optimality criteria, a handy characterization of the extremal ("adversarial") set families, and structure theorems for the case $s = 1$ that greatly limit the search space. Then we demonstrate these tools on the smallest r, where we also make some structural observations like symmetry breaking. This work should provide a stepping stone towards a full analysis of the greedy algorithm. It shows that we can "linearize" and therefore manage the continuous aspects. In further research it remains to understand the combinatorial side.

2 The basic concepts

We may identify a bit string p with the set $\{i : p_i = 1\}$ if this causes no confusion. Reading 0 and 1 as Boolean values we define $p \wedge q$ by $(p \wedge q)_i := p_i \wedge q_i$ for all i. Similarly we define $p \vee q$. We say that p and q intersect if $p \wedge q \neq o$ (the zero string). We write $p \subset q$ if $p_i \leq q_i$ for all i. Recall $n = |C|$. We define the *size* of a subset of C as its normalized cardinality: a set with xn elements has size x. Below we define the coverage of a set family $\mathcal{F}$ as the fraction of pairs in C covered by $\mathcal{F}$. But as we are concerned with fixed r and large n, we can neglect lower-order terms. This will greatly simplify our calculations.

Definition 1 A family $\mathcal{F}$ of r subsets $C_1, \ldots, C_r$ that induces a partitioning of C divides C into 2^r (possibly empty) *classes* $C(p)$, where each $p = p_1 \ldots p_r$ is a string of r bits, and $C(p) = \{v \in C | \forall i : v \in C_i \Leftrightarrow p_i = 1\}$. Let $x(p) = |C(p)|/|C|$ denote the size of $C(p)$. The pair coverage, simply *coverage*, of $\mathcal{F}$ is defined as

$$\pi(\mathcal{F}) := \sum_{p|\ p \neq o} x(p)^2 + 2 \left(\sum_{\{p,q\}|\ p \neq q,\ p \wedge q \neq o} x(p) \cdot x(q) \right)$$

The factor 2 appears, since the second sum is taken over unordered pairs. It is not hard to see that $\pi(\mathcal{F})$ is indeed the fraction of pairs in C covered by $\mathcal{F}$, subject to an $O(1/n)$ term for each summand. Thus, for fixed r this deviation vanishes as n grows. Clearly, many summands in the equation in Definition 1 can be zero. With some abuse of notation we also call $x(p)$ the size of p, and we call the string p and the class $C(p)$ *positive* if $x(p) > 0$.

Recall that r and s are fixed. Throughout the paper let c denote the optimal coverage that can be achieved by a sub-family of r sets from a given set family, and let g denote the coverage of the family obtained by the first s greedy steps. We call $1 - g/c$ the *missing fraction* of covered pairs.

Notation c and g can be similarly defined for PARTIAL SET COVER, with elements rather than pairs. Then the mentioned result [5,8] for PARTIAL SET COVER can be rephrased as follows: Every greedy step reduces the missing fraction by a factor $1-1/r$ or better, this guarantee is optimal, and this result does not depend on c. We also remark that the optimality proof is merely based on the pigeonhole principle and induction on s, see [5]. It turns out that, for PARTIAL SET COVER OF PAIRS, it is much more intricate to figure out the optimal missing fractions. The intuitive reason is that pairs impose additional graph-theoretic structure on the problem. However, any improved result for the first greedy steps (which might be easier to analyze than the general problem) yields immediate progress for general s as well: Suppose that we can prove, for some s_0, that the missing fraction is strictly smaller than $(1 - 1/r)^{s_0}$. Then we can still apply the general result for PARTIAL SET COVER to the subsequent greedy steps that reduce the missing fraction by a factor $1 - 1/r$ each. Thus we obtain improved (although not the best possible) results also for any $s > s_0$ greedy steps.

As we want to calculate approximation guarantees, we imagine an adversary that wants to present a set family that fools the greedy algorithm:

Definition 2 Let r, s, t be any fixed integers with $r, s \leq t$, and let c be any fixed real number with $0 < c \leq 1$. We call a set family $\mathcal{G}$ with t sets an *adversarial* family if some sub-family $\mathcal{F} \subset \mathcal{G}$ of r sets has the optimal coverage c among all sub-families of r sets, and the coverage g of the greedy solution with s sets from $\mathcal{G}$ is as small as possible, for the given r, s, t, and c.

Lemma 1 *For any fixed numbers r, s and c there exists an adversarial family with $t \leq r + s - 1$ sets that minimizes g, across all possible values of t.*

Proof In a given set family $\mathcal{G}$ we keep only the sets of $\mathcal{F}$ (defined above) and those selected by the greedy algorithm; any further sets can be removed without changing c and g. These are at most $r + s$ sets. Moreover, the last greedy set is always from $\mathcal{F}$, since otherwise we could remove it from $\mathcal{G}$ without making c worse, whereas g can only decrease. Thus, at most $r + s - 1$ sets remain. □

By *extending* a set we mean that we add more elements to it. Trivially, if a set in a family is extended, the coverage of this family can only increase. Similarly, *shrinking* a set means to subtract some elements from it. The benefit of the following lemma is that it will be easier to work with fixed g.

Lemma 2 *For any fixed integers r, s, t and $c < 1$, a set family $\mathcal{G}$ is adversarial if and only if no change of the class sizes $x(p)$ can both increase the optimal coverage $\pi(\mathcal{F})$ to some value larger than c and preserve the greedy coverage g.*

Proof If we can obtain an optimal coverage $c' > c$ and keep the same greedy coverage g, we can afterwards shrink all sets by multiplying all class sizes $x(p)$, $p \neq o$, with a common factor below 1, such that we get back the optimal coverage c and obtain some greedy coverage $g' < g$, thus $\mathcal{G}$ was not adversarial.

Conversely, if $\mathcal{G}$ is not adversarial, we can change the class sizes to obtain a family with optimal coverage c but some greedy coverage $g' < g$. Next, we can always extend some set to increase the optimal coverage to some $c' > c$. (Take any uncovered pair $\{u, v\}$ and extend some set C_i with $u \in C_i$ to include v, too.) Further extensions of sets can only increase the optimal coverage further, and the greedy coverage will eventually reach g again. This last conclusion requires some care. It is not obvious that the greedy coverage is monotone when sets are extended (because the greedy algorithm may then switch to other sets), however, g' changes continuously with the class sizes and will eventually be 1, hence we reach a situation where $g' = g$. □

The effect of an infinitesimal change of a class size $x(p)$ on the coverage of a set family is given by the partial derivative. Since the coverage is a polynomial of degree 2, the derivatives are just linear functions in the variables $x(p)$. It will be convenient to express this observation as follows: Any change of some $x(p)$ to $x(p) + h$ with an infinitesimal $h > 0$ (a "change $x(p) + h$" for short) adds $\frac{\partial \pi(\mathcal{F})}{\partial x(p)} \cdot h$ to the coverage. Clearly, for every bit string p we have:

$$\frac{\partial \pi(\mathcal{F})}{\partial x(p)} = 2x(p) + 2\left(\sum_{q \mid p \neq q,\, p \wedge q \neq o} x(q)\right) = 2\left(\sum_{q \mid p \wedge q \neq o} x(q)\right)$$

Lemma 3 *If infinitesimal changes $x(p) + h(p)$ are applied to all strings p (under the obvious constraint $\sum_p h(p) = 0$), then $\pi(\mathcal{F})$ changes by*

$$2\sum_p h(p) \sum_{q|\, p\wedge q\neq o} x(q) = 2\sum_q \left(\sum_{p|\, p\wedge q\neq o} h(p) \right) x(q)$$

Lemma 3 follows instantly from the previous equation. We refer to the sum in parantheses as the *coefficient* of $x(q)$.

3 The first greedy step

Now we provide some more specialized tools for the first greedy step, that is, for the case $s = 1$. As argued earlier, results on the first greedy step imply already general approximation guarantess (for any s) for Partial Set Cover of Pairs which are better than those inherited from Partial Set Cover.

Definition 3 Consider a change $h(p)$ applied to the sizes $x(p)$ of the classes induced by a set family. The change is called *homogeneous* if:

1. The change is non-zero: $\exists p : h(p) \neq 0$.
2. No cell size becomes negative: $\forall p : x(p) + h(p) \geq 0$.
3. The total sum of changes is zero: $\sum_p h(p) = 0$.
4. All set sizes are preserved: $\forall i : \sum_{p|\, p_i=1} h(p) = 0$.

Theorem 1 *For any fixed r and $s = 1$, and for every c, there exists an adversarial family of coverage c that (i) consists of only r sets, which (ii) have equal size. Furthermore, a family satisfying (i) and (ii) is adversarial if and only if no homogeneous change increases the coverage.*

Proof Property (i) follows from Lemma 1, and (ii) is trivially achieved by extending the non-maximum sets. Next, the changes that preserve both g and property (ii) are exactly the homogeneous changes. Hence the claimed equivalence holds, as a special case of Lemma 2. □

Now our strategy for calculating approximation guaranteess for the first greedy set can be outlined as follows. Once we manage to characterize the structure of the adversarial families, it will be straightforward to express g as a function of c, or vice versa. Theorem 1 is used to narrow down the possible adversarial families. Due to Theorem 1 it suffices to identify homogeneous changes that raise c, and the increase is computed as in Lemma 3. We consider the simplest homogeneous changes, affecting the smallest number of classes:

Definition 4 Let a, b, p, q be any four distinct strings with $x(a) > 0$, $x(b) > 0$, $p \wedge q = a \wedge b$, and $p \vee q = a \vee b$. Then we call $x(a) - h$, $x(b) - h$, $x(p) + h$, $x(q) + h$ a *quartet change*. A quartet change with $a \subset b$ is a *rhombus change*.

Lemma 4 *Every quartet change is homogeneous, moreover, every rhombus change preserves or increases the coverage.*

Proof Examining the four possible combinations of p_i, q_i, a_i, b_i for each i we see that the size of C_i is unchanged, hence every quartet change is homogeneous. Suppose that, additionally, $a \subset b$. Then any bit string t intersects the following of p, q, a, b: either none, or all four, or b and p, or b and q, or b and p and q. The coefficient of $x(t)$ in Lemma 3 is positive in the last case, and zero in all other cases. Hence the coverage can only grow. □

We will use Lemma 4 as follows. Every pair of strings a, b with $x(a) > 0$ and $x(b) > 0$ gives rise to quartet changes, for any p and q as specified in the Lemma. Since, by Theorem 1, the change of the coverage must not be positive, Lemma 3 implies several linear inequalities with class sizes as variables. By contraposition, any adversarial family must satisfy $x(a) = 0$ or $x(b) = 0$ or all these linear inequalities. This greatly restricts possible adversarial families. (Of course, similar reasoning applies to homegeneous changes in general.) However, we also stress that this does not make the characterization of adversarial families straightforward. Due to the disjunctions above, we cannot simply apply Farkas' lemma for the solvability of systems of linear equations.

We define the (Hamming) *weight* of a bit string as the number of 1s, and the (Hamming) *distance* of two bit strings as the number of bits in which they differ. With some abuse of notation, the weight of the class $C(p)$ means the weight w of string p, and we also call $C(p)$ a *w-class.* (Do not confuse size and weight.) The next small lemma about coefficients (see Lemma 3) saves some recurring calculations; later on we will not explicitly mention when it is used.

Lemma 5 *In a homogeneous change, the coefficient of $x(t)$ for any 1-class $C(t)$ is zero. In a quartet change, the coefficient of $x(t)$ is nonzero if and only if t intersects exactly three of a, b, p, q.*

Proof If t has weight 1, let i be the unique index with $t_i = 1$. The strings p that intersect t are exactly those with $p_i = 1$. Since $|C_i| = \sum_{p|\, p_i=1} x(p)$ is not changed, we have $\sum_{p|\, p_i=1} h(p) = 0$, which is the first assertion. The assertion on quartet changes follows from this observation: If t intersects some of the quartet strings at some position j, it must intersect a second one with an opposite change, since $|C_j|$ is preserved. □

More substantial is the following theorem that extends Theorem 1. Using the fact that rhombus changes cannot lower the coverage, we confine the positive classes to some narrow "stripe" in the partial order of bit strings:

Theorem 2 *For any fixed r and $s = 1$, and for every coverage c, there exists an adversarial family that consists of only r sets of equal size, and where $x(a) = 0$ or $x(b) = 0$ holds for every pair of strings a, b with $a \subset b$ and distance at least 2.*

Proof We start from an adversarial family as in Theorem 1. Let a and b be any positive strings with $a \subset b$ and distance at least 2. We can assume that a has minimum weight among all positive strings, since otherwise we could replace a with some positive $a' \subset a$, obtaining a pair of strings a' and b with the mentioned properties. Similarly, we can assume that b has maximum weight among all positive strings. Since a and b have distance at least 2, there exist strings p and q with $a = p \wedge q$ and $b = p \vee q$. We

do a rhombus change on a, b, p, q. By Lemma 4 this does not affect the set sizes and can only increase c. Since the family was already adversarial, actually the coverage is preserved, and we reach $x(a) = 0$ or $x(b) = 0$. Altogether we get rid of some positive classes with minimum or maximum weight. By iterating this step we eventually loose all pairs of strings with a, b as specified above, and the assertion is proved. □

In the following we demonstrate the application of these tools. We start from Theorem 2 and then infer the existence of adversarial families with specific class sizes. We focus on $s = 1$ and the smallest r, but we consider the whole range of c. Remember that c and g denote the optimal and greedy coverage, respectively, and that g is the square of the (common) size of the sets, if $g = 1$. In the quartet changes we always use $h > 0$.

4 Case $r = 2$ and $s = 1$

We have $x(10) = x(01)$ due to the equal set sizes, and $x(11) = 0$ or $x(00) = 0$. Hence there exists an adversarial family that consists of either two disjoint sets of size $\sqrt{g} \leq \frac{1}{2}$, or of two sets of size $\sqrt{g} > \frac{1}{2}$ whose union contains all elements. Straightforward calculation yields $g = (1 - \sqrt{\frac{1-c}{2}})^2$, which grows from $\frac{1}{4}$ to 1, when c grows from $\frac{1}{2}$ to 1. Thus the missing fraction decreases from $\frac{1}{2}$ to 0, whereas for PARTIAL SET COVER it would constantly be $\frac{1}{2}$. In the subsequent cases we will omit this final step of explicit calculations, and we only identify the adversarial families.

5 Case $r = 3$ and $s = 1$

If $x(100) > 0$ and $x(011) > 0$, then the quartet change $x(100) - h$, $x(011) - h$, $x(110) + h, x(001) + h$ raises c by $(x(011) - x(110)) \cdot 2h$, thus $x(110) \geq x(011) > 0$. Since $|C_1| = |C_3|$, we also get $x(100) + x(110) = x(001) + x(011)$, from which $x(001) \geq x(100) > 0$ follows. Thus we can do the opposite quartet change as well, such that $x(110) = x(011)$ and $x(100) = x(010)$ follow. Since this reasoning also applies to all symmetric cases, all 2-classes have the same size, and so have all 1-classes.

The quartet change above is not applicable if $x(p) = 0$ or $x(q) = 0$ holds for all complementary pairs p, q (that is, $p \wedge q = 000$, $p \vee q = 111$). Assume that some empty classes in two complementary pairs have different weights, say $x(100) = x(110) = 0$. By $|C_1| = |C_3|$ this also means $x(001) = x(011) = 0$. But now $|C_2| > 0$ and equality of set sizes implies $x(010) > 0$ and $x(101) > 0$, a contradiction. Thus, either all 1-classes or all 2-classes are empty. By equality of set sizes again, all 2-classes have the same size, and so have all 1-classes.

If $x(000) > 0$ then all other positive classes have weight 1, thus the three sets are disjoint. If $x(111) > 0$ then all other positive classes have weight 2, that is, they pairwise intersect, hence $c = 1$. But then the change $x(110) + h, x(101) + h, x(011) + h$, $x(111) - 3h$ shrinks the sets while c remains 1, thus the family was not adversarial. It remains the case when $x(000) = x(111) = 0$. Let x_1 and x_2 denote the (common)

size of the 1-classes and 2-classes, respectively. Then $c = 3x_1^2 + 9x_2^2 + 12x_1x_2$ and $g = x_1^2 + 4x_2^2 + 4x_1x_2$, respectively, thus $g = \frac{1}{3}c + x_2^2$. For $\frac{1}{3} \le c \le 1$ we find that g grows from $\frac{1}{3}$ to $\frac{4}{9}$, and we have characterized the adversarial families.

6 Case $r = 4$ and $s = 1$

In order to better distinguish the quartet changes we denote them by the weights of the positive strings involved. Quartet changes and their implications are described with the understanding that, of course, they also apply in all symmetric cases, under permutations of the r sets.

6.1 The (1,2) quartet change and the case of positive 1-classes

If both $x(1000) > 0$ and $x(0110) > 0$, then the $(1, 2)$ quartet change $x(1000) - h$, $x(0110) - h, x(1100) + h, x(0010) + h$ increases c by $(x(0111) + x(0110) - x(1100) - x(1101)) \cdot 2h$, thus we obtain $x(1100) + x(1101) \ge x(0110) + x(0111) > 0$. Since $x(1000) > 0$ implies that $x(1101) = 0$, this simplifies to $x(1100) \ge x(0110) + x(0111) > 0$. If additionally $x(0010) > 0$, then $x(0111) = 0$, furthermore we can apply the opposite quartet change which yields $x(1100) = x(0110)$. This proves the following implications:

Suppose that $x(1000) > 0$ and $x(0010) > 0$. Then we have $x(1100) = x(0110)$ and, by symmetry, also $x(1001) = x(0011)$. The existence of at least two positive 1-classes also rules out any positive 3-classes. Since all set sizes are equal, we further conclude $x(0100) = x(0001)$. Since $|C_2| = |C_4|$, this further means $x(1100) + x(0110) = x(1001) + x(0011)$, from which we can conlcude that $x(1100) = x(0110) = x(1001) = x(0011) =: w$.

If also $x(0100) = x(0001) > 0$, it follows in the same way that all 2-classes have equal size, and so have all 1-classes. Consider the other case when $x(0100) = x(0001) = 0$. Then, since the set sizes are equal, we get $x(0101) > w \ge 0$. But now the $(1, 2)$-quartet change above, with positions 3 and 4 swapped, enforces $w \ge x(0101)$, a contradiction. This shows: If at least two 1-classes are positive, then all classes of equal weight have equal size.

Finally suppose that exactly one 1-class is positive, say $x(1000) > 0$. Recall that $x(0110) = 0$, or an $(1, 2)$ quartet change yields $x(1100) \ge x(0110) + x(0111)$. In either case we get $x(1100) \ge x(0110)$, and similarly in all six symmetric cases. By adding the resulting six inequalities we obtain $x(1100) + x(1010) + x(1001) \ge x(0110) + x(0101) + x(0011)$. Since $3|C_1| = |C_2| + |C_3| + |C_4|$, we get that $3x(1000) + 2x(1100) + 2x(1010) + 2x(1001) = 2x(0110) + 2x(0101) + 2x(0011) + 3x(0111)$, thus $x(0111) \ge x(1000) > 0$. The $(1, 3)$ quartet change $x(1000) - h$, $x(0111) - h$, $x(1100) + h, x(0011) + h$ increases c by $(x(0111) + x(0110) + x(0101) - x(1100)) \cdot 2h$, thus $x(0111) + x(0110) + x(0101) \le x(1100)$. By adding the three symmetric inequalities of this form, and another obvious step, we obtain $3x(0111) + 2x(0110) + 2x(0101) + 2x(0011) \le x(1100) + x(1010) + x(1001)$, $3x(1000) + 2x(1100) + 2x(1010) + 2x(1001) \le x(1100) + x(1010) + x(1001)$, a contradiction. Altogether we see that the classes of equal weight also have equal size, unless all 1-classes are empty. Let x_1 and x_2 denote the (common) size of all 1-classes and 2-classes, respectively.

Then $c = 4x_1^2 + 30x_2^2 + 24x_1x_2$ and $g = x_1^2 + 9x_2^2 + 6x_1x_2$, thus $g = \frac{1}{4}c + \frac{3}{2}x_2^2$, which grows from 0.25 to 0.2916. This characterizes the adversarial families for $\frac{1}{4} \leq c \leq \frac{5}{6}$.

6.2 The case that all 1-classes are empty

If $x(1110) > 0$ and $x(0011) > 0$, then the $(2, 3)$ quartet change specified by $x(1110) - h$, $x(0011) - h$, $x(1010) + h$, $x(0111) + h$ increases the coverage c by $(x(1100) - x(0101)) \cdot 2h$, thus $x(1100) \leq x(0101)$. By symmetry this also yields $x(1100) \leq x(1001)$. Furthermore, if $x(0101) > 0$, then a similar $(2, 3)$ quartet change yields $x(1010) \leq x(0011)$ and $x(1010) \leq x(1001)$.

We will also use the $(2, 2)$ quartet change: If $x(1100) > 0$ and $x(0011) > 0$, then $x(1100) - h$, $x(0110) + h$, $x(0011) - h$, $x(1001) + h$ increases the coverage c by $(x(1100) + x(0011) - x(0110) - x(1001)) \cdot 2h$, thus it holds $x(1100) + x(0011) \leq x(0110) + x(1001)$.

6.2.1 Subcase: All 2-classes are positive

Suppose that two 3-classes are positive, too, say $x(1110) > 0$ and $x(0111) > 0$. By applying the inequalites from $(2, 3)$ quartet changes we see $x(1010) = x(0011)$ and $x(1100) = x(0101)$. Now a new phenomenon arises: The superposition of two $(2, 3)$ quartet changes $x(1110) - 2h$, $x(0101) - h$, $x(0011) - h$, $x(1100) + h$, $x(1010) + h$, $x(0111) + 2h$ increases c by $(x(1100) + x(1010) - x(0101) - x(0011))) \cdot 2h$. That is, the variables that increase/decrease are on the positive/negative side, thus c will further increase by this symmetry breaking. Thus, at most one 3-class is positive, say $x(1110) \geq 0$. *Notably, classes of equal weight are no longer of equal size.*

Since $|C_i| = |C_4|$, for $i = 1, 2, 3$, we have $x(1110) + x(1100) + x(1010) = x(0101) + x(0011)$, and similarly in the symmetric cases. These three equations imply that $x(1001) - x(0110)$, $x(0101) - x(1010)$, $x(0011) - x(1100)$ are equal. Since the sums and differences, respectively, of the sizes of complementary 2-classes are equal, we conclude in the case $x := x(1110) \geq 0$ that $x(1001) = x(0101) = x(0011) =: y$ and $x(1100) = x(1010) = x(0110) =: z$. Observe $g = 9y^2$, $c = 1 - 6yz$, $x + 3y + 3z = 1$, and $2y = x + 2z$, where the last equation comes from the equality of set sizes. Straightforward algebra yields $g = \frac{1}{4}c + \frac{1}{24} + \frac{1}{6}x + \frac{1}{96}x^2$ for $0 \leq x \leq \frac{2}{5}$.

6.2.2 Subcase: Exactly one 2-class is empty

Say $x(1100) = 0$. If both $x(1110) > 0$ and $x(1101) > 0$, we can again raise c by $(2, 3)$-changes, thus only one of these classes is positive, say only $x(1110) > 0$. Furthermore, we have $x(0011) \geq x(1010) + x(0101)$ and $x(0011) \geq x(1001) + x(0110)$, since otherwise some $(2, 2)$ quartet change could raise c. As $|C_1| + |C_2| = |C_3| + |C_4|$ and $|C_3| = |C_4|$, we have $x(1110) \geq 2x(0011) \geq x(1010) + x(0101) + x(1001) + x(0110)$ and $x(1110) + x(1010) + x(0110) = x(1001) + x(0101)$. This yields $x(1010) + x(0110) = 0$, a contradiction, hence this case is impossible.

6.2.3 Subcase: At least two 2-classes are empty

If only two complementary 2-classes are empty, then a $(2, 2)$ quartet change can increase c. Hence two incident 2-classes must be empty, say $x(1100) = x(1010) = 0$. Now, if both $x(1001) > 0$ and $x(0110) > 0$, then both $x(0101) \geq x(1001) + x(0110)$

and $x(0011) \geq x(1001) + x(0110) > 0$ must hold to prevent $(2, 2)$ quartet changes from raising c. Since $|C_1| = |C_2|$, we conclude $x(1001) + x(1011) = x(0110) + x(0101) + x(0111) \geq x(1001) + 2x(0110) > x(1001)$, thus $x(1011) > 0$. Since $x(0110) > 0$, a $(2, 3)$ quartet change can be avoided only by $x(1001) \leq x(1100) = 0$, a contradiction. It follows $x(1001) = 0$ or $x(0110) = 0$. With the remaining possible positive classes the coverage is 1, and due to earlier work [3] on the special case $c = 1$, the only adversarial family in this case is given by $x(1110) = \frac{2}{5}$ and $x(1001) = x(0101) = x(0011) = \frac{1}{5}$.

Acknowledgements The information retrieval motivation of this work comes from the project "Data-driven secure business intelligence", supported by Grant IIS11-0089 from the Swedish Foundation for Strategic Research (SSF).

References

1. Chvatal, V.: A greedy heuristic for the set-covering problem. Math. Oper. Res. **4**, 233–235 (1979)
2. Cygan, M., Kratsch, S., Pilipczuk, M., Pilipczuk, M., Wahlström, W.: Clique cover and graph separation: new incompressibility results. ACM Trans. Comput. Theory **6**, 6 (2014)
3. Damaschke, P.: Pairs covered by a sequence of sets. In: Kosowski, A., Walukiewicz, I. (eds.) FCT 2015. LNCS, vol. 9210, pp. 214–226. Springer, Heidelberg (2015)
4. Edwards, K., Griffiths, S., Kennedy, W.S.: Partial interval set cover–trade-offs between scalability and optimality. In: Raghavendra, P., Raskhodnikova, S., Jansen, K., Rolim, J.D.P. (eds.) APPROX-RANDOM 2013, LNCS, vol. 8096, pp. 110–125. Springer, Heidelberg (2013)
5. Elomaa, T., Kujala, J.: Covering analysis of the greedy algorithm for partial cover. In: Elomaa, T., Mannila, H., Orponen, P. (eds.) Algorithms and Applications. Essays dedicated to Esko Ukkonen on the occasion of his 60th birthday, LNCS, vol. 6060, pp. 102–113. Springer, Heidelberg (2010)
6. Feige, U.: A threshold of in n for approximating set cover. J. ACM **45**, 634–652 (1998)
7. Gramm, J., Guo, J., Hüffner, F., Niedermeier, R.: Data reduction and exact algorithms for clique cover. ACM J. Exper. Algorithmics **13**, 2 (2008)
8. Hochbaum, D.S., Pathria, A.: Analysis of the greedy approach of maximum k-coverage. Naval Res. Q. **45**, 615–627 (1998)
9. Kneis, J., Langer, A., Rossmanith, P.: Improved upper bounds for partial vertex cover. In: Broersma, H., Erlebach, T., Friedetzky, T., Paulusma, D. (eds.) WG 2008 LNCS, vol. 5344, pp. 240–251. Springer, Heidelberg (2008)
10. Niedermeier, R.: Invitation to Fixed-Parameter Algorithms. Oxford lecture series in math. and its appl. Oxford University Press, Oxford (2006)